AD-A231 277

COPY

(19)

REPORT DOCUMENTATION PAGE

Form Approved
OPM No. 0704-0188

1. AGENCY USE ONLY *(Leave Blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	August 1990	Final

4. TITLE AND SUBTITLE

Methods to Assess the Utility of Proxies

5. FUNDING NUMBERS

C - N00014-87-C-0001

PE - 65153M

PR - C0031

6. AUTHOR(S)

Neil B. Carey

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Center for Naval Analyses
4401 Ford Avenue
Alexandria, Virginia 22302-0268

8. PERFORMING ORGANIZATION REPORT NUMBER

CRM 89-311

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Commanding General
Marine Corps Combat Development Command (WF 13F)
Studies and Analyses Branch
Quantico, Virginia 22134

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This research memorandum reviews methods for quantifying the tradeoffs between using proxy (i.e., surrogate) measures of job performance versus the established benchmark criterion of hands-on performance tests. Such analytical methods must be sensitive to the intended application of the proxy. Two applications that require precise performance information are examined for equivalence of outcomes when proxies are used as opposed to hands-on tests.

14. SUBJECT TERMS

Aptitude, Aptitude tests, Enlisted personnel, Infantry, Infantrymen, Marine Corps personnel, Performance (human), Performance tests, Personnel selection, Proficiency, Qualifications, Reliablity, Statistical analysis, Test methods, Validation, Work measurement

15. NUMBER OF PAGES

48

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
CPR	CPR	CPR	SAR

CNA CENTER FOR NAVAL ANALYSES
A Division of Hudson Institute 4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

26 September 1990

MEMORANDUM FOR DISTRIBUTION LIST

Subj: Center for Naval Analyses Research Memorandum 89-311

Encl: (1) CNA Research Memorandum 89-311, *Methods to Assess the Utility of Proxies*, by Neil B. Carey, August 1990

1. Enclosure (1) is forwarded as a matter of possible interest.

2. This research memorandum reviews methods for quantifying the tradeoffs between using proxy measures of job performance versus the established benchmark criterion of hands-on performance tests. Such analytical methods must be sensitive to the intended application of the proxy. Two applications that require precise performance information are examined for equivalence of outcomes when proxies are used as opposed to hands-on tests.

Lewis R. Cabe
Director
Manpower and Training Program

Distribution List:
Reverse page

Subj: Center for Naval Analyses Research Memorandum 89-311

Distribution List

SNDL

45A2	CG I MEF
45A2	CG II MEF
45A2	CG III MEF
45B	CG FIRST MARDIV
45B	CG SECOND MARDIV
A1	ASSTSECNAV MRA
A1	DASN MANPOWER (2 copies)
A2A	CNR
A6	HQMC M'R &RA
	Attn: Code M
	Attn: Code MR
	Attn: Code MP
	Attn: Code MM
	Attn: Code MA (3 copies)
A6	CG MCRDAC, Washington
A6	HQMC AVN
FF38	USNA
	Attn: Nimitz Library
FF42	NAVPGSCOL
FF44	NAVWARCOL
	Attn: E-111
FJA1	COMNAVMILPERSCOM
FJB1	COMNAVCRUITCOM
FJA13	NAVPERSRANDCEN
	Attn: Technical Director (Code 01)
	Attn: Technical Library
	Attn: Director, Manpower Systems (Code 61)
	Attn: Director, Personnel Systems (Code 62)
FT1	CNET
V8	CG MCRD Parris Island
V8	CG MCRD San Diego
V12	MCCDC
	Attn: Studies and Analyses Branch
	Attn: Director, Warfighting Center
	Attn: Warfighting Center, MAGTF Proponency and Requirements Branch (2 copies)
	Attn: Director, Training and Education Center
V12	CG MCRDAC, Quantico

OPNAV

OP-01
OP-11
OP-12
OP-13

OTHER

Defense Advisory Committee on Military Personnel Testing (8 copies)
Joint Service Job Performance Measurement Working Group (13 copies)

CRM 89-311 / August 1990

Methods to Assess the Utility of Proxies

Neil B. Carey

Force Structure and Acquisition Division

CNA *A Division of Hudson Institute*

CENTER FOR NAVAL ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

ABSTRACT

This research memorandum reviews methods for quantifying the tradeoffs between using proxy (i.e., surrogate) measures of job performance versus the established benchmark criterion of hands-on performance tests. Such analytical methods must be sensitive to the intended application of the proxy. Two applications that require precise performance information are examined for equivalence of outcomes when proxies are used as opposed to hands-on tests.

EXECUTIVE SUMMARY

It is a continual challenge for the armed services to determine the qualifications of applicants, to assess the effects of training, and to document troops' state of readiness. Each of these challenges requires the use of reliable, empirically validated performance measures. The congressionally mandated Job Performance Measurement (JPM) project is a joint-service effort to obtain such performance information. The project focuses on hands-on performance tests (HOPTs) as the benchmark measure of job performance.

Despite the many advantages of hands-on tests, there are also several drawbacks. First, HOPTs are expensive to develop and administer. These tests also tend to expend costly resources (such as electrical parts or ammunition), may endanger personnel or equipment (e.g., working with land mines), or require use of scarce equipment (such as operational aircraft) so as to limit other training opportunities. In addition, test security is difficult to maintain for HOPTs because they are individually administered by trained scorers.

This paper explores methods to analyze the usefulness of proxies (i.e., surrogates) as substitute performance measures for HOPTs. A companion paper, *Assessment of Surrogates for Hands-On Tests: Selection Standards and Training Needs* (CRM 90-47), uses the methods proposed in this paper to analyze proxies of infantry job performance. A surrogate is a test that resembles a HOPT for a particular purpose so closely that it is considered "equivalent." There are three important criteria for a proxy to be equivalent to a HOPT: comparability of reliability, validity, and decision outcomes.

When evaluating potential proxies, it is important to realize that no proxy is equivalent for all purposes. This paper illustrates methods for evaluating six surrogates (job-knowledge tests, training grades, proficiency marks, video marksmanship trials, supervisor ratings, and conduct marks) as to their suitability for two particular uses: (1) setting classification standards,[1] and (2) diagnosing

1. *Classification standards* are requirements for assignment to occupational specialties (MOSs) within a service branch. In the Marine Corps, these standands are determined on the basis of a composite of subtests in the Armed Services Vocational Aptitude Battery (ASVAB). Depending on the Marine Corps MOS, the composite might be General Technical (GT), Clerical/Administrative (CL), Mechanical Maintenance (MM), or Electrical (EL). Before assessing whether an applicant meets classification standards for particular specialties, the services determine whether the candidate meets "selection" or "enlistment" standards on the basis of mental aptitude, educational level, physical fitness, moral character, age, and citizenship. The Armed Forces Qualification Test (AFQT) is the primary indicator of enlistment aptitude and recruit quality for setting selection standards.

training needs. Table I summarizes the differences in methods for analyzing proxies. Analytical methods that are appropriate for setting classification standards are not necessarily appropriate for diagnosing training needs, and vice versa.

Table I. Evaluating proxies for setting standards versus diagnosing training needs

Setting standards	Diagnosing training needs
1. Analyze scores at the lower part of the distribution, depending on reasonable baserate[a] assumptions.	1. Analyze scores at *all* parts of the distribution to see whether there are differences in the training needs of troops of different aptitudes.
2. Analyze the overall composite, and validity by *occupational field.*	2. Analyze duty area scores separately by MOS.
3. Illustrate the usefulness of the proxy, given different baserate and selection ratio[b] assumptions.	3. Illustrate the usefulness of the proxy, given different duty area assumptions.

a. *Baserate* means the proportion of examinees who would become competent Marines if every examinee were accepted and placed in a MOS. If 60 out of the 100 examinees would be competent if all examinees were accepted, the baserate would be 60 percent.
b. *Selection ratio* means the number of persons selected divided by the number of applicants.

One method, first proposed by Maier and Mayberry,[1] meets the criteria in table I for setting classification standards. In this paper, Maier and Mayberry's procedure for using the "10-percent rule"[2] was used to set hypothetical classification standards using each proxy.

Next, methods to determine whether a proxy would be useful for diagnosing training needs were analyzed. Based on this analysis, it was concluded that profiles of duty area scores for HOPTs and proxies should be compared.

1. CNA Research Memorandum 89-9, *Evaluating Minimum Aptitude Standards*, by Milton H. Maier and Paul W. Mayberry, July 1989.
2. The 10-percent rule states that a standard should result in no more than a 10-percent failure rate of trainees from basic training school.

This paper addresses how to analyze the usefulness of proxies for hands-on tests. It concludes that the following steps should be taken to evaluate proxies (figure I):

1. Determine how the prospective proxy will be used. Plan an analysis based on the expected use of the surrogate.

2. If the proxy is to be used for setting classification standards, compute reliability and validity coefficients across all subtests. Compute a composite standard using the 10-percent rule based on the present criterion, and compare this with the composite standard using the prospective surrogate. Determine whether the composite standard would vary by base.

3. If the prospective surrogate is to be used to diagnose training needs, plot duty-area strengths and weaknesses based on surrogate scores and HOPT scores. If the pattern of duty-area weaknesses for HOPTs and proxy match, then the prospective surrogate will result in comparable decision outcomes. Otherwise, further analyses are needed to determine the reasons for incompatible results (e.g., fallibility of testing mode for concept being measured).

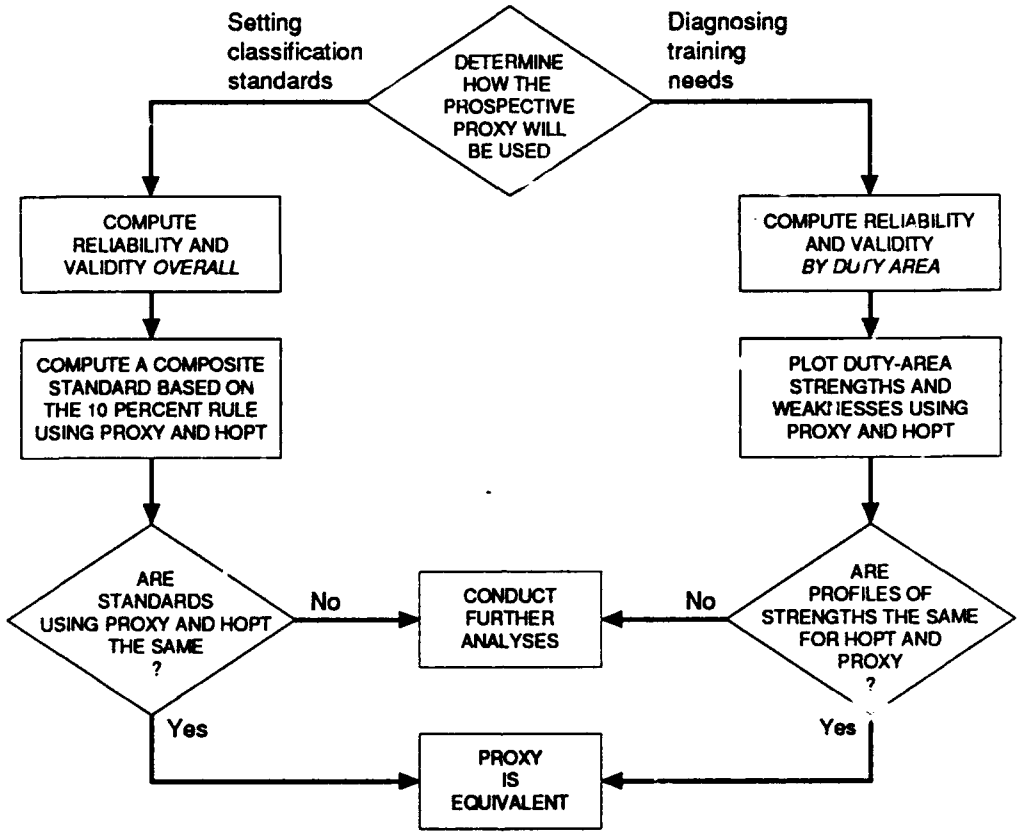


Figure I. Steps for analyzing the equivalence of proxies

CONTENTS

	Page
Illustrations	xi
Tables	xi
Introduction	1
Review of Previous Research on Proxies	1
A Model	4
Evaluating the Usefulness of Proxies to Set Classification Standards	6
Results	10
Critique	12
Assessing the Usefulness of Proxies to Diagnose Training Needs	18
Conclusions	21
References	23
Appendix A: Illustration and Definitions of Commonly Used Terms Associated With Standard Setting	A-1–A-2
Appendix B: Cross-Sample Prediction Frequencies for the 75-Percent Selection Ratio	B-1–B-3
Appendix C: Cross-Sample Prediction Frequencies for the 25-Percent Selection Ratio	C-1–C-3

ILLUSTRATIONS

		Page
1	Hypothesized Relationships Among Cost, Completeness, and Purity of Criterion and Proxy Variables	4
2	Relationship Between Hands-On and Predicted Hands-On Performance (Cross Model)	9
3	Profiles of Training Needs Using Hands-On Performance Test and Job-Knowledge Test	21

TABLES

		Page
1	ASVAB Qualification Standards for High School Graduates	3
2	Evaluating Proxies for Setting Standards Versus Diagnosing Training Needs	5
3	Summary of Results for Six Potential Surrogates	11
4	Summary of Results Taking Top 75 Percent, 50 Percent, and 25 Percent of HOPT and Surrogate Scorers	13
5	Principal Components Analysis of Ten ASVAB Subtest Scores of Riflemen	14
6	Summary of Results for Seven Potential Surrogates	14
7	Summary of Results for Four Small Samples from MOS 0311	16
8	GT Standards Using Different Proxies for HOPTs	19
9	Stability of GT Standard and Validity by Base Using 10th-Percentile Cutoff	20

INTRODUCTION

This paper reviews several methods for analyzing the usefulness of proxy (i.e., "surrogate") measures of enlisted job performance. Many methods of analysis are misleading or do not clearly communicate the tradeoffs involved in using proxy measures for setting classification standards[1] or diagnosing training needs. This paper illustrates methods that provide more complete information concerning proxies, and concludes with an analysis plan for the Marine Corps Job Performance Measurement (JPM) infantry data.

Performance information is important to the Marines because of the continual challenge to determine standards, assess the effects of training, and document Marines' state of readiness. Each of these challenges ideally requires the use of reliable, empirically validated performance measures.

For these purposes, the most valid method to assess job performance is by hands-on testing [4]. Hands-on performance tests (HOPTs), however, have several disadvantages: they are costly, can be dangerous to personnel and equipment, and can require the transfer of resources from unit training to individual testing. Furthermore, test security and consistency for hands-on tests are more difficult to maintain because these tests are administered individually.

To avoid the problems involved in using HOPTs, the services could use a proxy (i.e., surrogate), which is a test that resembles a HOPT for a particular purpose so closely that it is considered "equivalent." But how does one know whether a proposed proxy would be equivalent? This paper reviews methods for evaluating the tradeoffs among potential proxies.

REVIEW OF PREVIOUS RESEARCH ON PROXIES

Gottfredson [5] has reviewed ways to analyze potential proxies for the National Academy of Science committee that oversees the work of the Joint-Service JPM Project. Her paper emphasizes that an analysis of potential surrogates must begin

1. *Classification standards* are requirements for assignment to occupational specialties (MOSs) within a service branch. In the Marine Corps, these standards are determined on the basis of a composite of subtests in the Armed Services Vocational Aptitude Battery (ASVAB). Depending on the Marine Corps MOS, the classification composite might be General Technical (GT), Clerical/Administrative (CL), Mechanical Maintenance (MM), or Electrical (EL) [1, 2]. Before assessing whether an applicant meets classification standards for particular specialties, the services determine whether the candidate meets "selection" or "enlistment" standards on the basis of mental aptitude, education level, physical fitness, moral character, age, and citizenship [3]. The Armed Forces Qualification Test (AFQT) is the primary indicator of enlistment aptitude and recruit quality for setting selection standards.

with knowledge of the purpose for the proxy. A measure that provides a valid substitute for one purpose may be inappropriate for another.

Allred [6] has written a paper for the same committee, reviewing alternatives to the correlation coefficient for describing the relationship between two variables. Allred's paper illustrates the importance of knowing what part of the performance distribution is critical for the particular use of the test, since a test cannot be equally useful at detecting differences in performance in all parts of the distribution. For example, a test might be efficient in detecting differences among low-aptitude examinees, but unreliable at distinguishing among those with higher aptitudes. If this "ceiling effect" occurred, all examinees at the upper end of the distribution might get the highest possible scores on the test.

Together, the Gottfredson and Allred papers suggest that different analyses are required, depending on the purpose for which a proxy will be used. A proxy that is useful for setting classification standards may not be useful for diagnosing training needs, and vice versa.

Many proxies have been tried in past research. May [7] has developed a method, based on the professional judgments of Marine Corps officers, to translate proficiency and fitness marks into measures of enlisted Marines' relative value to the service. The rescaled proficiency and fitness marks were used to calculate performance differences between high school graduates and nongraduates.

Hiatt [8] has analyzed school and field proficiency marks as measures of job performance because they offer a readily available proxy for hands-on performance without the added expense of new data collection. Also, proficiency marks indicate whether a Marine *will* do a job, whereas hands-on performance tests measure only whether a Marine *can* do a job.

Hiatt separated proficiency marks into two categories: those given at the end of formal school training (school ratings) and those awarded once a Marine is working in the field (field proficiency or PRO marks). Hiatt found a stronger relationship between ASVAB scores and school ratings than between field proficiency ratings and these scores. The field proficiency of high school graduates was rated consistently higher than their nongraduate counterparts, but PRO and conduct (CON) ratings were subject to a *halo effect* in which the PRO and CON marks were strongly correlated. This finding suggests that raters are not as proficient in detecting differences between attitude and proficiency as they are in detecting overall performance levels.

Hiatt also searched for trends in ratings to suggest an enlistment standard based on PRO and CON marks. Higher ASVAB scores were associated with higher

ratings, but ratings did not suggest an enlistment standard because there was no particular ASVAB score at which the ratings tended to level off. (Although it is not necessary for ratings to level off in order to set an ASVAB standard, if ratings leveled off above a certain score, that would indicate use of that score as a selection cutoff). Hiatt concluded that "ratings do not appear to be suitable, by themselves, for setting enlistment standards. They may, however, be useful as part of a composite measure of performance that could be used to set enlistment standards" (p. vii).

Maier and Hiatt [9] analyzed ASVAB, HOPT, training grades, and job-knowledge test scores for ground radio repair personnel, automotive mechanic personnel, and infantry riflemen. From the correlation of written tests and HOPT, Maier and Hiatt concluded that in the two technical skill categories (radio repair and automotive mechanic), the written tests and training grades "show promise as substitutes for the hands-on tests. For the infantry rifleman skill, the written test shows promise as a substitute for the hands-on test, but because of the lower correlation with the hands-on test, training grades show less promise" (p. iv).

Maier and Hiatt next evaluated the ASVAB qualification standards that would result from using hands-on job performance as the criterion for validating ASVAB. They assumed that varying percentages of the population would be satisfactory radio repairers, automotive mechanics, and infantry riflemen, respectively. They also assumed that the Marine Corps could tolerate a failure rate of 10 percent. Using a combination of hands-on and written proficiency tests as criteria, they found qualification standards as listed in table 1. It is notable that, given their assumptions, Maier and Hiatt found a correspondence of existing ASVAB standards based on training grades and those based on hands-on job performance and written proficiency tests.

Table 1. ASVAB qualification standards for high school graduates (from Maier and Hiatt [9])

	Qualification standards[a]	
Skill	Existing	Derived
Ground radio repair	115	115
Automotive mechanic	90	95
Infantry rifleman	80	85

a. Existing standards are for high school graduates; derived standards were estimated by Maier and Hiatt's study [9].

A MODEL

Figure 1 illustrates the ways that proxies might vary. The first dimension, purity, refers to the degree to which the measure is objective, and the degree to which it taps intended attributes without measuring unintended characteristics. It is hypothesized that subjective measures such as field proficiency marks, supervisor ratings, and (to a certain extent) grade-point averages would be less pure criteria than measures such as a HOPT or job-knowledge test, which would be more objective. Another dimension, *completeness*, is the degree to which a test *completely* measures job performance. Notice that field proficiency marks are complete because they refer not only to the "can do" part of proficiency, but also to the "will do" part. The third criterion, cost, refers to the approximate expense of each measure.

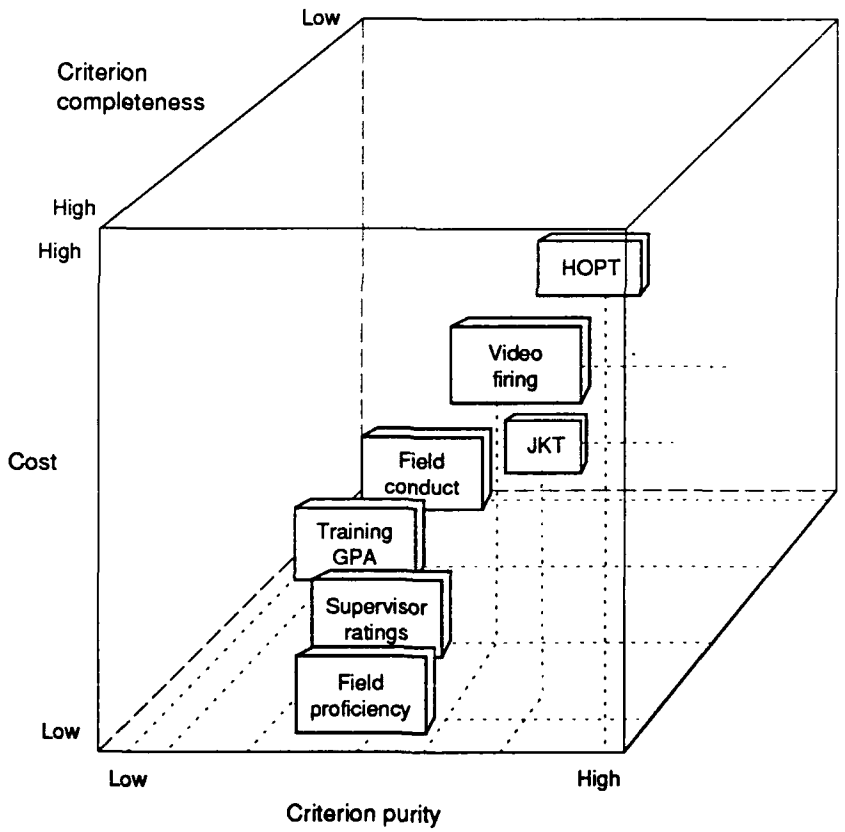


Figure 1. Hypothesized relationships among cost, completeness, and purity of criterion and proxy variables

Table 2 summarizes the implications of the Gottfredson and Allred papers. These two papers suggest that different kinds of analyses should be done, depending

on whether a measure is being evaluated for setting classification standards or for diagnosing training needs. Since setting standards is generally a matter of screening out those who would not be competent performers, the focus in standard setting is on the lower part of the distribution. In contrast, tests developed for the diagnosis of training needs should measure ability in all parts of the distribution, since it is important to find out how all trainees have benefited from learning opportunities.

Table 2. Evaluating proxies for setting standards versus diagnosing training needs

Setting standards	Diagnosing training needs
1. Analyze scores at the lower part of the distribution, depending on reasonable baserate assumptions.	1. Analyze scores at *all* parts of the distribution to see whether there are differences in the training needs of troops of different aptitudes.
2. Analyze the overall composite, and validity by *occupational field.*	2. Analyze duty area scores separately by MOS.
3. Illustrate the usefulness of the proxy, given different baserate and selection ratio assumptions.	3. Illustrate the usefulness of the proxy, given different duty area assumptions.

Classification tests are developed to place a prospective Marine into a particular field, so the policy-maker does not require a detailed synopsis of particular strengths and weaknesses. In contrast, a proxy used to diagnose training needs must provide detailed knowledge of the duty areas in which examinees excel or fail. Training information should be specific to MOS, whereas for classification it is necessary to widen the focus to a particular occupational field (e.g., the infantry field rather than solely the rifleman specialty within the field).

Finally, an analysis of a classification test must consider the baserate and selection ratio for which the test will be used.[1] Work by Taylor and Russell [10],

1. *Selection ratio* means the number of people selected divided by the number of applicants. *Baserate* means the proportion of examinees who would become competent employees if every examinee were accepted (i.e., no test was used). If 60 out of the 100 examinees would be competent if all applicants were accepted, the baserate would be 60 percent. (See appendix A for an illustration.)

shows that as the selection ratio decreases, the advantage of a more valid test is more apparent.[1] For example, for a selection ratio of .95, the difference in the percentage of correctly chosen personnel between a test with a .10 validity and a .60 validity is merely 4 percent, whereas for a selection ratio of .10, the difference is almost 58 percent.

Similarly, as the baserate increases, the advantage of a more valid test is less apparent. For a baserate of .5, a test with a validity of .6 and a selection ratio of .1 will identify acceptable personnel 58 percent more often than will a test with a validity of .1. The advantage is merely 7.5 percent if the baserate increases to .9.

In contrast to a surrogate for standard setting, a proxy for diagnosing training standards should be evaluated by the degree to which the test validly assesses areas of relative strength and weakness. For example, it would be important to know whether infantrymen need more training in land navigation tasks, throwing hand grenades, or first aid.

The model and studies just reviewed [1 through 10] suggest that a number of methods could be used to evaluate surrogates for HOPTs, depending on the purpose for which the proxy would be used. We will now look at methods for evaluating proxies for (1) setting classification standards, and (2) diagnosing training needs.

EVALUATING THE USEFULNESS OF PROXIES TO SET CLASSIFICATION STANDARDS

Recent research by Hanser [11] presents a method to evaluate proxies for setting classification standards that meet the criteria listed in table 1. This method, sometimes called cross-validated regression, involves determining whether use of a proxy would result in a significantly different number of correct[2] classification decisions. By this reasoning, if use of a proxy yields approximately the same proportion of correct selection decisions, then the proxy is "equivalent" for some classification purposes. With Hanser's method, the crucial element is the proportion of correct selections; different individuals may be accepted even when the proportion of correct decisions is the same. This section critiques the use of this method for evaluating proxies.

---

1. Taylor-Russell tables assume bivariate normality of the data.
2. A "correct" decision means either (1) accepting someone who later meets or surpasses a given performance standard on the criteria, or (2) rejecting one who later would have failed to meet the performance standard.

To illustrate the cross-validated regression method, this author used the technique on the six Marine Corps proxies. The following procedures were used:

1. The sample of 1,804 cases was randomly assigned to one of two groups.

2. The hands-on scores were regressed on the ten ASVAB subtests, by group. This resulted in a hands-on regression equation for each group, as follows:

Group 1	Group 2
$\hat{HOCORE}$[1] = 5.45 + 0.17GS + 0.13AR	$\hat{HOCORE}$ = 13.72 + 0.12GS + 0.12AR
− 0.04WK + 0.05PC + 0NO	− 0.08WK + 0.05PC − 0.01NO
− 0.01CS + 0.28AS + 0.12MK	+ 0.01CS + 0.22AS + 0.10MK
+ 0.11MC + 0.15EI	+ 0.16MC + 0.11EI

3. Each potential surrogate (e.g., JKTCORE, PRO marks, CON marks, GPA, video firing, supervisor ratings) was regressed on the ten ASVAB subtests, by group. This resulted in a series of regression equations for each group, e.g., for JKTCORE:

Group 1	Group 2
$\hat{JKTCORE}$[2] = − 17.11 + 0.24GS + 0.22AR	$\hat{JKTCORE}$ = − 14.23 + 0.12GS + 0.26AR
+ 0WK + 0.09PC + 0.05NO	+ 0WK + 0.27PC + 0.03NO
+ 0.10CS + 0.16AS + 0.17MK	+ 0.06CS + 0.12AS + 0.01MK
+ 0.06MC + 0.12EI	+ 0.14MC + 0.14EI

4. The regression coefficients from the opposite group (step 2) were used to develop a predicted HOPT score for each individual.

5. The regression coefficients from the opposite group (step 3) were used to develop a predicted surrogate score for each individual, e.g., for $\hat{JKTCORE}$.

6. Actual HOPT performance was plotted against the predicted HOPT performance (i.e., $\hat{HOCORE}$) and against predicted surrogate performance with cutoffs at the 25th and 75th percentiles. There is no special significance to these cutoff points, although the 25th percentile is similar to that used for

1. $\hat{HOCORE}$ is the predicted core hands-on performance for each individual, based on regressing hands-on performance on ASVAB. The "^" character above HOCORE is used to indicate that these are *predicted* scores. It is not *actual* HOPT performance.
2. $\hat{JKTCORE}$ is the predicted job-knowledge test performance, based on the regression of JKT performance on ASVAB. It is not *actual* JKT performance.

some infantry specialties and the 75th percentile is similar to the standards for some technical specialties. Two-by-two tables (see Appendix B) showing the number of true positives, true negatives, false positives, and false negatives with each surrogate and actual HOPT performance were developed. The percentage of correct decisions and the percentage of competent people accepted using each surrogate were then compared.

Figure 2 illustrates how two-by-two tables were developed. In panel A, the solid vertical line shows the 25th percentile cutoff, while the corresponding horizontal line illustrates the 25th percentile HOPT standard. Those points falling to the right of the vertical line have "passed" on the ASVAB composite, and those above the horizontal line have demonstrated acceptable HOPT performance. In panel B, the dashed lines show the cutoff and standards for the 75th percentile. Those in the upper right corner (defined by the vertical cutoff and corresponding horizontal standard) are "true positive"—those who would be accepted by the ASVAB composite and who at least equalled satisfactory performance on the actual criterion. Those in the lower left corner (defined by the cutoff and standard) are "true negatives"—those rejected by the ASVAB composite and who failed to meet the standard on the hands-on test. The lower right corner of the vertical and horizontal lines corresponds to "false positive" (those who pass the ASVAB composite but fail to meet the hands-on standard), and the upper left hand corner corresponds to "false negatives" (those who are rejected by the ASVAB composite but who would have passed the hands-on standard).

Figure 2 also illustrates how each surrogate will be evaluated. In panel A, for example, if we wanted to accept the best 75 percent in terms of actual hands-on performance, we should take those whose scores are above the horizontal solid line (i.e., everyone above the 25th percentile). Since we can only predict their performance given the surrogate model, we will accept those whose predicted scores fall to the right of the solid vertical line. In doing so, we have mistakenly taken those in the lower right quadrant (false positives) and omitted those in the upper left quadrant (false negatives). The dashed lines in panel B can be interpreted the same way, except that these are for the higher standard of choosing the top 25 percent of the sample.

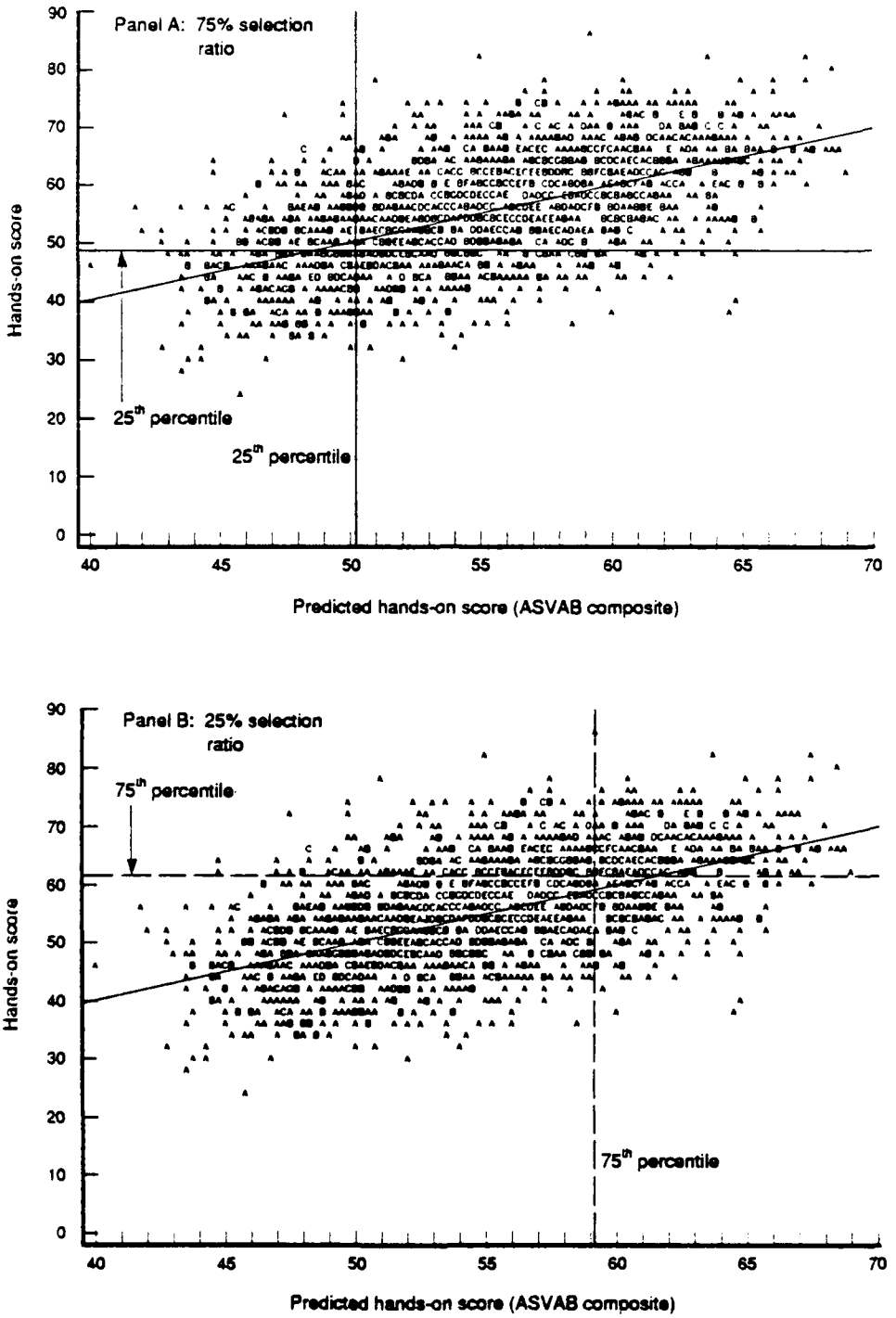


Figure 2. Relationship between hands-on and predicted hands-on performance (Cross model)

RESULTS

Appendixes B and C show the number of correct decisions versus incorrect decisions for each surrogate at 75-percent and 25-percent selection ratios, respectively. These analyses, summarized in table 3, show the percentage of correct selections,[1] additional percentage of correct selections, percentage of total possible added, average HOPT performance, and the standard deviation for each surrogate and the HOPT.[2] Note that for a 75-percent selection ratio, possible values range from random selection, with an average of 75-percent correct selections and hands-on performance of 55.58, up to an average score of 59.33 if selection were perfect (i.e., 100-percent correct selections).

For a 75-percent selection ratio, note in table 3 that field conduct ratings add only 6.1 percent correct solutions, whereas HOPT and infantry school GPA add 10.2 percent and 9.9 percent correct selections, respectively. Similarly, use of the HOPT or infantry school GPA would add a total of 2.08 or 2.00 points to the average HOPT performance, respectively. These numbers correspond to approximately .21 standard deviation improvement over random selection.

For a 25-percent selection ratio, using field proficiency ratings would add 22.8 percent to the percentage of correct selection decisions, while the HOPT and infantry school GPA would add 28.6 percent and 25.9 percent, respectively. Note that using the job-knowledge test adds almost six points, or .63 standard deviation, to the average hands-on performance above what would occur if random selection were used.

The table 3 column labeled "percentage of total possible added" indicates how much a surrogate adds to the percentage of correct selections between random and perfect selections. For the 75-percent selection ratio, there is a total of 25 percent possible to be added to the percentage of correct selection decisions. For example, the infantry school GPA adds 9.9 percent to the percentage of correct selections out of a possible 25-percent improvement over the expected percentage of correct selections using random selection. This corresponds to a 9.9/25 = 39.6 percent of the total possible improvement.

The numbers in table 3 indicate significant differences between different measures. For the 75-percent selection ratio, infantry school GPA adds 41 percent as much to the possible improvement as does a field proficiency rating (39.6 percent

1. For the rest of the paper "selection" will be used to mean acceptance or classification into an occupational field—specifically, the infantry (0300).
2. Appendix A defines and illustrates terms that will be used throughout the rest of the paper.

versus 28.0 percent). For the 25-percent selection ratio, in contrast, infantry school GPA adds less to the percentage of total possible added (34.5 percent versus 30.4 percent), but makes a slightly larger impact on the average hands-on performance compared to proficiency ratings (a difference of 0.95 for the 25-percent ratio; a difference of only .62 for the 75-percent selection ratio). These numbers contradict Hanser's finding (using Army JPM data) that predicted job-knowledge test scores would result in more correct decisions than would HOPT scores [11]. In other words, table 3 indicates that some proxies are more useful than others, in contradiction to Hanser's findings.

Table 3. Summary of results for six potential surrogates

	Percentage of correct selections	Percentage correct above random	Percentage of total possible added	Average hands-on performance	Standard deviation
75-percent selection ratio					
Perfect selection	100.0%	25.0%	100.0%	59.33	6.72
Hands-on	85.2	10.2	40.8	57.66	8.55
Core job-knowledge test	83.7	8.7	34.8	57.50	8.74
Infantry school GPA	84.9	9.9	39.6	57.58	8.56
Video firing	84.5	9.5	38.0	57.56	8.61
Supervisor rating	82.3	7.3	29.2	56.99	9.02
Field proficiency ratings	82.0	7.0	28.0	56.96	8.89
Field conduct ratings	81.1	6.1	24.4	56.88	9.17
Random selection baseline	75.0	0.0	0.0	55.58	9.45
25-percent selection ratio					
Perfect selection	100.0%	75.0%	100.0%	66.77	4.08
Hands-on	53.6	28.6	38.1	61.72	7.56
Core job-knowledge test	50.9	25.9	34.5	61.35	7.87
Infantry school GPA	50.9	25.9	34.5	61.19	7.88
Video firing	49.7	24.7	32.9	60.91	7.91
Supervisor rating	47.0	22.0	29.3	60.31	8.24
Field proficiency ratings	47.8	22.8	30.4	60.24	8.32
Field conduct ratings	42.2	17.2	22.9	59.00	9.29
Random selection baseline	25.0	0.0	0.0	55.58	9.45

NOTE: "Perfect selection" refers to taking those in the top 75 percent or 25 percent of HOPT scores. Figures for the random selection baseline are based on the estimate of the mean and standard deviation (s.d.) derived from JPM data (for average performance and s.d.), or are expected values over repeated sampling (for percentages of correct selections). The percentage correct for any particular random sample could vary considerably from the figures shown above.

CRITIQUE

There are weaknesses in using cross-validated regression as a method to evaluate surrogates. This method only creates an ASVAB composite, which is quite different from setting an ASVAB standard. Creating a standard requires confidence in the criterion, and a method to determine what level of the criterion is minimally acceptable. The series of regression coefficients extracted from cross-validation has no inherent meaning upon which to judge what is minimally acceptable. A job expert could not judge whether a certain set of regression weights "makes sense," or whether one set of ten scores is better than another. In contrast, a job expert could make judgments about the acceptability of a score on a hands-on performance test or job-knowledge test.

If an ASVAB standard remains the same, exactly the same people will be selected. This is not the case if, as Hanser's method would suggest, a different surrogate is used just because it results in a similar percentage of correct decisions using cross-validated regression. Although the percentage of correct selections is approximately the same across different measures for cross-validated regression (table 3), the same people are not necessarily being selected if a different proxy is used. Ignoring this fact could result in unduly minimizing important differences between surrogates.

The Marines were rank-ordered on HOPTs and the six surrogates, and the top 25 percent were compared on each measure, the job-knowledge test identified 60.6 percent of the top HOPT scorers, while the surrogates other than training school GPA identified less than 45 percent. GPA identified only 37 percent of top HOPT performers. Table 4 shows that the difference between using the job-knowledge test and using proficiency ratings is considerable at the 50-percent and 25-percent selection ratios. Improvements for the 50- and 25-percent selection ratio are 16 percent and 36 percent, respectively.

Two final flaws of cross-validated regression for evaluation of surrogates are that (1) other methods can extract regression coefficients more easily, without reference to surrogates, and (2) sampling error can overly influence outcomes concerning which surrogate is "best."

Table 4. Summary of results taking top 75 percent, 50 percent, and 25 percent of HOPT and surrogate scorers

	Percentage of correct selections		
	75-percent selection ratio	50-percent selection ratio	25-percent selection ratio
Job-knowledge test	83.4%	70.3%	60.6%
Field proficiency ratings	81.2	60.6	44.7
Video firing test	83.4	62.8	44.3
Field conduct scores	79.1	59.6	40.4
Supervisor ratings	79.7	58.4	38.6
Training school grade-point average (GPA)	82.4	59.9	37.0

The point that methods other than cross-validated regression can extract regression coefficients more easily requires some explanation. Cross-validated regression improves on random selection by creating a composite of ASVAB scores that captures general ability. Composites developed with cross-validated ASVAB coefficients to predict HOPT and job-knowledge test scores are the most successful in predicting HOPT performance because HOPT and job knowledge are the best measures of general ability. However, another set of regression coefficients that capture general ability could be derived by performing principal components analysis of the ten ASVAB subtests, without reference to the relationship between ASVAB and a criterion. Principal components analysis extracts the dependence of scores in a set of correlation data [12], simplifying its structure and separating error from components of ability measured. Therefore, principal components can extract the dependencies of different ASVAB subtests and create a simplified description of the relationships among subtests.

To illustrate the use of principal components, riflemen (MOS 0311) were randomly divided into two groups and separately analyzed using principal components. The two-factor solutions for both groups are shown in table 5. The vector described as factor 1 captures general ability for power tests, whereas factor 2 captures ability on the speeded tests NO and CS.

Table 6[1] shows the results when loadings for factor 1 were used in place of regression coefficients. Figures below the dashed line are the seven "surrogates" to be compared to the random-selection baseline. As can be seen, use of principal components of the ASVAB

1. Table 6 with MOS 0311 is used here rather than table 3 (which contained all MOS) because table 7 will separate these data and compare the effects of sampling variability. It is better to use a single MOS for this analysis so that the effects of multiple MOSs do not cloud the later discussion of sampling variability.

results in second-highest average performance and percentage of correct selections for surrogates for a 75-percent selection ratio, and the third-highest results among surrogates for a 25-percent selection ratio. In both cases, the principal-components surrogate performs better than supervisor ratings, proficiency ratings, and field conduct marks.

Table 5. Principal components analysis of ten ASVAB subtest scores of riflemen (MOS 0311)

ASVAB subtest	Group 1		Group 2	
	Factor 1	Factor 2	Factor 1	Factor 2
GS	0.80126	–0.11514	0.80565	–0.19383
AR	0.65762	0.42344	0.65446	0.38750
WK	0.75205	–0.22721	0.76850	–0.25071
PC	0.62094	–0.02251	0.60984	0.00705
NO	–0.14926	0.81082	–0.04236	0.83207
CS	0.06886	0.73584	0.16845	0.71208
AS	0.70385	–0.18962	0.69229	–0.19847
MK	0.66994	0.39460	0.64551	0.42363
MC	0.79155	0.03002	0.79343	0.03570
EI	0.78568	–0.10903	0.74243	–0.21304

Table 6. Summary of results for seven potential surrogates (for entire MOS 0311, $n = 1{,}020$)

Measure	75-percent selection ratio		25-percent selection ratio	
	Percentage of correct selections	Average HOPT performance	Percentage of correct selections	Average HOPT performance
Perfect selection	100.0%	59.16	100.0%	66.69
Predicted hands-on	85.2	57.45	56.6	61.99
Principal components	84.6	57.38	53.7	61.57
Predicted core job-knowledge test	84.1	57.33	56.4	61.98
Predicted supervisor rating	81.0	56.50	39.9	58.95
Predicted field proficiency rating	82.9	56.92	53.3	61.18
Predicted grade point average (GPA)	84.0	57.25	56.6	62.00
Predicted video score	85.0	57.43	52.5	61.19
Predicted conduct score	79.9	56.32	50.0	60.36
Random selection baseline	75.0	55.24	25.0	55.24

The point that sampling error can overly influence judgments about which surrogate is "best" is illustrated in tables 6 and 7. For table 7, the sample of 1,020 riflemen that was illustrated in table 6 was randomly divided into four samples of 255. Each of the smaller samples was analyzed using principal components and cross-validated regression. The resulting percentage of correct selections and average HOPT performance was then calculated for each sample. Table 6 showed that cross-validation using HOPT resulted in the highest percentage of correct selections and the highest average HOPT performance among those selected for the entire sample of 1,020, but table 7 demonstrates that conclusions drawn from smaller samples would vary considerably. For the 75-percent selection ratio, the principal-components surrogate results in the highest percentage of correct selections twice, the job-knowledge test (core JKT) does so once, and the HOPT does once. For the 25-percent selection ratio, four different methods result in the highest percentage of correct selections, depending on the sample chosen.

In summary, the cross-validated regression method has four shortcomings as a way to judge surrogates:

- First, this method asks the wrong question. The issue is whether standards would be the same with different surrogates, because the same people will be chosen if standards stay the same. Cross-validated regression results in a composite that has no inherent meaning and, hence, should not be used to set a standard.

- Second, the cross-validated regression method masks the fact that different people are selected if different surrogates are used. Two surrogates can appear to be the same, even though it is not a matter of indifference to the individual which surrogate is used.

- Third, this is a method that could be simplified and often improved by using principal-components analysis.

- Finally, the conclusions of this method can be overly affected by sampling error.

The method used by Maier and Mayberry [13] called the 10-percent rule solves the problems associated with the cross-validated regression method. The rationale for the Maier and Mayberry method is that historically the Marine Corps has set standards based on an expectation of a maximum 10-percent failure rate of trainees from basic training school. The method is therefore based on the assumption that an ASVAB standard should result in no more than 10 percent of the eligible target population failing in hands-on tests at the end of training. The target population was obtained from the 1980 Youth Population study. The 1980 Youth Population

Table 7. Summary of results for four small samples from MOS 0311 (N = 255)

75-Percent selection ratio	Percentage of Correct Selections					AVE HOPT Performance				
	Gp A	Gp B	Gp C	Gp D	Ave	Gp A	Gp B	Gp C	Gp D	Ave
Perfect selection	100.0	100.0	100.0	100.0	100.0	59.13	58.89	58.83	59.39	59.06
Predicted hands-on (HOPT)	83.5	85.8	87.8	82.1	84.8	57.34	57.54	57.36	56.88	57.28
Principal components	85.2	85.2	86.1	84.2	85.2	57.65	57.31	57.01	57.33	57.38
Predicted core JKT	83.1	86.3	85.6	80.9	84.0	57.20	57.51	57.22	56.40	57.08
Predicted supervisor rating	79.8	79.3	82.3	77.6	79.8	55.93	55.49	56.13	55.73	55.82
Predicted field proficiency	79.2	78.7	82.8	81.0	80.4	55.93	56.04	56.26	56.03	56.07
Predicted GPA	83.0	83.1	82.8	78.8	81.9	57.00	57.02	56.35	56.18	56.64
Predicted video firing	82.4	84.8	84.4	83.1	83.7	57.29	57.34	56.50	56.97	57.03
Predicted field conduct	76.0	76.0	78.9	79.9	77.7	55.60	54.81	55.22	55.76	55.35
Random selection baseline	75.0	75.0	75.0	75.0	75.0	55.22	55.22	55.22	55.22	55.22
25-Percent selection ratio	Percentage of Correct Selections					AVE HOPT Performance				
	Gp A	Gp B	Gp C	GP D	Ave	Gp A	Gp B	Gp C	Gp D	Ave
Perfect selection	100.0	100.0	100.0	100.0	100.0	66.84	66.27	65.61	66.84	66.39
Predicted hands-on	48.4	57.4	68.9	52.5	56.8	60.97	61.48	62.95	60.67	61.52
Principal components	48.4	50.8	59.0	57.4	53.9	61.02	60.82	61.61	61.13	61.14
Predicted core JKT	50.0	51.6	70.5	54.8	56.7	61.11	61.29	62.82	61.15	61.59
Predicted supervisor rating	34.4	34.4	40.0	36.1	36.2	57.21	56.41	57.45	57.89	57.24
Predicted field proficiency	45.9	32.8	52.5	41.0	43.0	59.18	56.87	59.41	58.82	58.57
Predicted GPA	46.8	54.8	54.1	51.6	51.8	60.94	61.03	60.75	60.61	60.83
Predicted video firing	55.7	48.4	44.3	57.4	51.4	62.02	60.19	59.03	61.67	60.73
Predicted field conduct	42.6	26.2	43.3	35.5	36.9	58.03	55.15	56.43	57.68	56.82
Random selection baseline	25.0	25.0	25.0	25.0	25.0	55.22	55.22	55.22	55.22	55.22

Note: The *n* for each group was 255. The random selection baseline percentage of correct selections is an expected value for repeated samples. Actual values would vary from sample to sample. The average performance is the estimate derived from the JPM data. Actual values would vary.

data[1] are consequently used to develop a standard that approximately 10 percent of the eligible male population would fail to meet.

To illustrate the 10-percent-rule method, the following steps were taken to develop infantry standards based on JPM data:

1. The sample raw correlation coefficients relating the General Technical (GT) composite to hands-on performance were corrected for multivariate restriction [14], so that these values approximated the correlation in the general population.

2. Since infantrymen are selected for their occupational field on the basis of the General Technical (GT) composite of ASVAB, the corrected correlation coefficients were used to get corrected estimates for each of the surrogates for the general population:

$$\text{Surrogate: } = B_{0_i} + B_{1_i} * GT + B_{2_i} * TIS + e_i$$

3. Because infantrymen are supposed to be competent at 24 months, each regression equation was then used to compute predicted values of performance on the surrogate for the 1980 Youth Population, based on their GT scores. Time-in-service (TIS) was set at a constant value of 24.

4. Since errors of prediction have been removed from the above regression equations, they must be added back by introducing a random component to each computed score. This was accomplished by creating a random normal deviate for each eligible male in the population, multiplying this by the standard error of the estimate (SEE) of the sample regression equation, and adding the resulting product to the predicted performance score.

5. The resulting values of performance on each proxy were ordered, and the proxy score corresponding to the 10th percentile was chosen.

6. The proxy score corresponding to the 10th percentile was then substituted in the left side of the regression equation in step 2. The value of four months was substituted for TIS, since in mobilization, infantrymen are expected to be proficient by the time they finish training school. The resulting equation was then solved for GT, which is the value that would predict mean

1. The 1980 Youth Population data provide a nationally representative sample of 18- to 23-year-old males and females who took the ASVAB. The population used here was restricted to males (since the focus is on combat specialties) and excluded persons of extremely low aptitude who are legally ineligible for service (called category V personnel).

performance at the 10th percentile on the proxy. This value is the computed cutoff for GT, using each measure as a substitute for HOPT scores.

7. If the GT standard computed for the proxy is nearly identical to that computed for the criterion, then the proxy is "equivalent" with respect to setting infantry classification standards. If, in addition, the GT standard computed using the proxy varies little by year or base, then the proxy should be considered as a substitute for the HOPT for the purpose of setting standards.

Table 8 shows the GT standards computed using the 10-percent rule. The table shows that although roughly equal numbers of correct decisions are made using the cross-validated regression method described previously, very different GT standards would be computed if most surrogates were used in place of the HOPT. Only the job-knowledge test comes close to the standard of 80 computed with the hands-on performance test [13]. In addition, this table illustrates that most surrogates, if used in place of hands-on performance, would result in a lowering of classification standards, primarily because these surrogates have lower validities than the HOPT or JKT. Once correlation coefficients have been standardized, they are proportional to the regression coefficients of the criterion on the predictors.

This table illustrates the usefulness of the 10-percent rule as a method to evaluate proxies, because if a proxy results in the same GT criterion, then *exactly the same people will be selected if the proxy is used in place of the criterion.* In this sense, the proxy is certainly "equivalent," not just in the proportion of correct decisions, but in who is selected.

Table 9 shows that the GT standard for all proxies except the job-knowledge test would change substantially by base. This finding suggests that proxies must be analyzed by base to determine their usefulness for setting standards. Differences in grading philosophy also have implications for using grade-point average in setting standards. These findings demonstrate that the job-knowledge test is the most useful proxy for setting standards.

ASSESSING THE USEFULNESS OF PROXIES TO DIAGNOSE TRAINING NEEDS

One method that may be used to assess the usefulness of a proxy for diagnosing training needs would be to determine whether the conclusions would vary if the proxy were used in place of the criterion. If the proxy and the HOPT result in similar conclusions about which duty areas may require more training, then the proxy is considered "equivalent" for the purposes of diagnosing training needs.

Table 8. GT standards using different proxies for HOPTs

Surrogate	Regression equation	GT Cutoff At 10th percentile	Population validity
Core JKT[a]	JKT = –13.5 + .53GT + .13TIS	81	.70
School of Infantry GPA[b]	GPA = 19.0 + .28GT + .08TIS	67	.44
Video marksmanship[c]	VIDEO = 110.5 + .79GT + .26TIS	60	.47
Field proficiency[d]	FPRO = 26.3 + .17GT + .21TIS	52	.37
Field conduct[e]	FCON = 34.2 + .12GT + .14TIS	19	.27
Supervisor ratings[f]	RATING = 74.5 + .18GT + .27TIS	8	.28

NOTE: Actual GT scores can be no lower than 40. The cutoffs computed for field conduct marks and supervisor ratings demonstrate how poorly these surrogates perform for setting standards.

a. Predicted job-knowledge test scores ranged from a low of 4 to a high of 85. The minimum score of 30 corresponded to a cumulative percentage of 11.4, and the SEE for the regression was 8.5.
b. Predicted grade-point averages across the two schools of infantry (Pendleton and Lejeune) ran from a low of 18 to a high of 88. The score of 38 corresponded to a cumulative percentage of 11.5, and the SEE for the regression was 9.3.
c. The predicted video marksmanship scores ranged from a low of 95 to a high of 320. The minimum performance of 159 corresponded to a cumulative percentage of 10.3, and the SEE of the regression was 30.2.
d. Predicted field proficiency scores ranged from a low of 17 to a high of 88. The score of 36 corresponded to a 10.1 cumulative percentage, and the SEE for the regression was 8.5.
e. Predicted conduct scores ranged from a low of 16 to a high of 87. The score of 37 corresponded to a cumulative percentage of 11.8, and the SEE for the regression was 8.5.
f. Predicted supervisor ratings ranged from a low of 41 to a high of 165. The minimum score of 77 corresponded to a cumulative percentage of 10.1. The SEE for the regression was 17.5.

Table 9. Stability of GT standard and validity by base using 10th-percentile cutoff

	GT Standard			GT Validity		
	Base A	Base B	Difference	Base A	Base B	Difference
Job-knowledge test	79	81	2	.74	.80	.06
Video marksmanship	66	54	12	.47	.43	.04
Field proficiency	43	57	14	.25	.35	.10
Grade-point average	73	57	16	.63	.42	.21
Field conduct	2	31	29	.17	.25	.18
Supervisor rating	0	29	29	.07	.26	.19

NOTE: Actual GT scores range from a low of 40. The computed standards for field conduct and supervisor ratings demonstrate how poorly these two surrogates perform for setting classification standards.

To accomplish this evaluation, mean scores over all MOSs (0311, 0331, 0341, 0351) for each of the 12 duty areas were standardized[1] to develop a profile of strengths and weaknesses. The profiles of each surrogate and the criterion test could then be compared to determine whether the two measures agree. If the pattern is the same for the criterion and HOPT, then the proxy is considered equivalent.

The only surrogate that provided detailed information down to the duty-area level was the job-knowledge test. Figure 3 shows a profile of strengths and weaknesses (standardized mean-duty area scores) based on the HOPT (solid line) and on the job-knowledge test (dotted line).[2] The 12 duty areas, from left to right, are communications (CM); first aid (FA); grenade launchers (GL); hand grenades (HG); light antitank weapons (LAW); land navigation (LN); nuclear, biological, and chemical defense (NBC); night vision (NV); squad automatic weapons (SAW); security and intelligence (SI); and tactical measures (TM). The figure shows some discrepancies between the conclusions indicated by the HOPT and the job-knowledge test. The HOPT indicates that first aid, hand grenades, and land navigation are duty areas that require more training, whereas the job-knowledge test indicates that first aid, communications, night vision, and tactical measures are areas of relative weakness. Job-knowledge test and HOPT results are particularly

1. Means for each duty area were standardized within each measurement mode (HOPT or JKT). In other words, each point represents the standard score $(X_i - \bar{X})$s.d), where $\bar{X}$ is the grand mean of all 12 duty area means.
2. The shaded area indicates where two-thirds of the points would be expected to fall by chance. Points falling outside the shaded area are not likely to occur by chance.

discrepant for hand grenades and night vision. In the case of hand grenades, troops apparently understand how to throw a hand grenade (as evidenced by the job-knowledge test), but they cannot throw one well in practice (as evidenced by the HOPT). The night vision area shows the opposite pattern: troops can use night vision equipment but are not proficient in answering questions on how to perform night vision procedures.

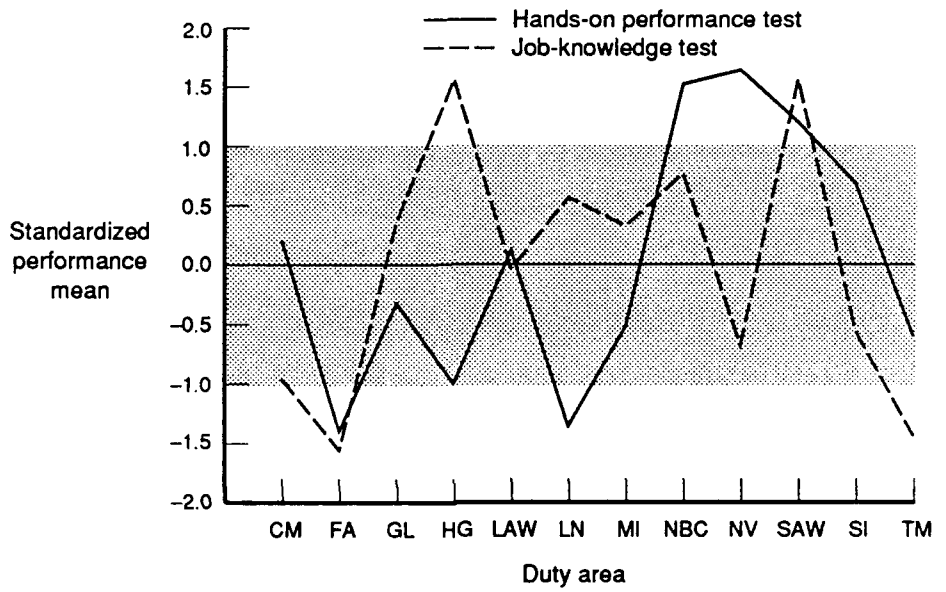


Figure 3. Profiles of training needs using hands-on performance test and job-knowledge test

CONCLUSIONS

In conclusion, different methods must be employed for evaluating proxies, depending on how these surrogates are to be used operationally. The following is a plan that would use the best of the methods demonstrated in this paper:

- Determine how the prospective proxy will be used. Plan an analysis based on the expected use of the surrogate.

- If the proxy is to be used for setting classification standards, compute reliability and validity coefficients across all subtests. Compute a composite standard using the 10-percent rule based on the present criterion, and compare this with the composite standard using the prospective surrogate. Determine whether the composite standard would vary by base.

- If the prospective surrogate is to be used to diagnose training needs, Compute reliability and validity coefficients by duty area. Plot duty-area strengths and weaknesses based on surrogate scores and HOPT scores. If the pattern of duty-area strengths for HOPT and proxy match, then the prospective surrogate will result in comparable decision outcomes. Otherwise, further analyses are needed to determine the reasons for incompatible results (e.g., fallibility of testing mode being measured).

REFERENCES

[1] CNA Research Memorandum 89-290, *Effect of the GT Composite Requirement on Qualification Rates*, by Neil B. Carey, Mar 1990 (27890290)[1]

[2] CNA Research Memorandum 90-47, *An Assessment of Surrogates for Hands-On Tests: Selection Standards and Training Needs*, by Neil B. Carey, Jul 1990 (27900047)

[3] Brian K. Waters, Janice H. Laurence, and Wayne J. Camara. *Personnel Enlistment and Classification Procedures in the U.S. Military*. Washington, D.C.: National Academy Press, 1987

[4] Committee on the Performance of Military Personnel, Commission on Behavioral and Social Sciences and Education, National Research Council. *Job Performance Measurement in the Military: Report of the Workshop*. National Research Council/National Academy of Sciences, Sep 1986

[5] Linda Gottfredson. *The Evaluation of Alternative Measures of Job Performance*. Paper prepared for the Committee on the Performance of Military Personnel, Commission on Behavioral and Social Sciences and Education, National Research Council/National Academy of Sciences, Sep 1986

[6] Linda J. Allred. *Alternatives to the Validity Coefficient for Reporting the Test-Criterion Relationship*. Paper prepared for the Committee on the Performance of Military Personnel, Commission on Behavioral and Social Sciences and Education, National Research Council/National Academy of Sciences, Washington, D.C., Jul 1986

[7] CNA Research Contribution 550, *The Translation of Supervisory Ratings into Measurements of Relative Value*, by Laurie J. May, Unclassified, Jul 1986 (02055000)

[8] CNA Research Contribution 537, *Supervisor Ratings Analysis*, by Catherine M. Hiatt, Unclassified, Feb 1986 (02053700)

[9] CNA Report 89, *An Evaluation of Using Job Performance Tests To Validate ASVAB Qualification Standards*, by Milton H. Maier and Catherine M. Hiatt, May 1984 (94008900)

1. The numbers in parentheses are CNA control numbers.

REFERENCES (Continued)

[10] H.C. Taylor and J.T. Russell. "The Relationship of Validity Coefficients to the Practical Effectiveness of Tests in Selection: Discussion and Tables," *Journal of Applied Psychology 23, (1939) 565-578*

[11] *Developing Performance Relevant Capability Indices: Examining Surrogate Measures.* Briefing to the Joint-Service Job Performance Measurement Working Group, by Lawrence M. Hanser, Aug 1989

[12] D.F. Morrison. *Multivariate Statistical Methods.* New York: McGraw-Hill, 1976

[13] CNA Research Memorandum 89-9, *Evaluating Minimum Aptitude Standards*, by Milton H. Maier and Paul W. Mayberry, Jul 1989 (27890009)

[14] Harold Gulliksen. *Theory of Mental Tests.* Hillsdale, NJ: Lawrence Erlbaum Associates, 1978

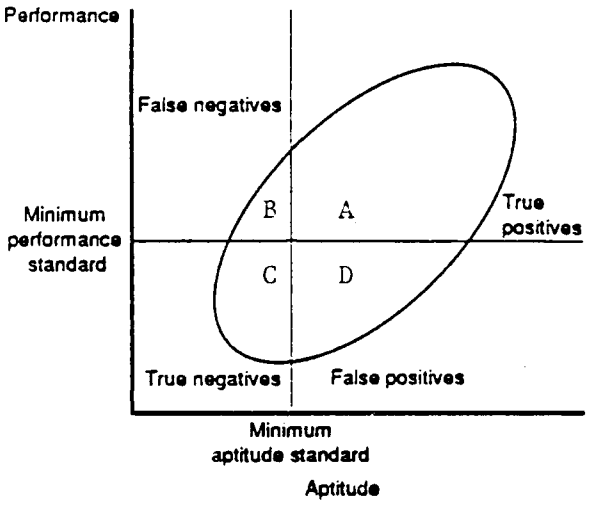
APPENDIX A

ILLUSTRATION AND DEFINITIONS OF COMMONLY USED TERMS ASSOCIATED WITH STANDARD SETTING

APPENDIX A

ILLUSTRATION AND DEFINITIONS OF COMMONLY USED TERMS ASSOCIATED WITH STANDARD SETTING

The following figure and definitions introduce the reader to elementary concepts of personnel decision-making.



Decision outcomes associated with minimum performance and aptitude standards

Baserate refers to the proportion of people who would be competent if all examinees were accepted. In the illustration, it refers to the proportion (A + B)/(A + B + C + D).

False negatives are those examinees who fail to meet the aptitude standard but who would have met the minimum performance standard.

False positives are those examinees who meet the aptitude standard but who do not meet the minimum performance standards.

Hit rate is the proportion of all classifications made correctly: (A + C)/(A + B + C + D).

Percentage of correct selections is the proportion of successes among those actually selected: A/(A + D).

Selection ratio refers to the number of persons selected divided by the number of applicants: (A + D)/(A + B + C + D).

True positives are those examinees who meet the aptitude standard and who meet the minimum performance standard.

True negatives are examinees who fail to meet the aptitude standard and who would not have met the minimum performance standard.

APPENDIX B

CROSS-SAMPLE PREDICTION FREQUENCIES FOR THE 75-PERCENT SELECTION RATIO

APPENDIX B

CROSS-SAMPLE PREDICTION FREQUENCIES FOR THE 75-PERCENT SELECTION RATIO

The following tables show the decision outcomes of using regression models based on a HOPT and six surrogates to predict actual hands-on performance of infantry tasks. Each table shows the result of using a 75-percent selection ratio cutoff using a particular model. To build the model, the entire sample of 1,804 was randomly split into two equal groups. Each model was then developed by separately regressing the two groups' criterion (HOPT) or surrogate scores (e.g., job-knowledge test) on the ten enlisted ASVAB subtests (GS, AR, WK, PC, NO, CS, AS, MK, MC, and EI). Regression coefficients from the opposite group were used to cross-validate each model. "Actual success" refers to those who scored in the top 75th percentile in the hands-on criterion, while "actual failure" refers to those who scored in the bottom 25 percent. "Predicted success" means obtaining a score in the top 75th percentile in the predicted criterion (HOPT or surrogate) from the cross-validated regression composite of 10 ASVAB subtests. "Hit rate" is the proportion of all classification decisions made correctly, whereas "percentage of correct selections" is limited to the proportion of successes among those who actually would be selected on the basis of the model.

Table B-1. Selection using HOPT model: 75-percent selection ratio

Actual HOPT performance percentile above 25	Predicted performance using HOPT model	
	Failure	Success
Success	201	1,158
Failure	244	201

Note: Hit rate = (244 + 1,158)/1804 = 77.7%
Percentage of correct selections = 1,158/1,359 = 85.2%

Table B-2. Selection using core job-knowledge test model: 75-percent selection ratio

Actual HOPT performance percentile above 25	Predicted performance using core job-knowledge test model	
	Failure	Success
Success	221	1,138
Failure	224	221

Note: Hit rate = (224 + 1,138)/1,804 = 75.5%
Percentage of correct selections = 1,138/1,359 = 83.7%

Table B-3. Selection using field proficiency model: 75-percent selection ratio

Actual HOPT performance percentile above 25	Predicted performance using field proficiency model	
	Failure	Success
Success	244	1,115
Failure	200	245

Note: Hit rate = (200 + 1,115)/1,804 = 72.9%
Percentage of correct selections = 1,115/1,360 = 82.0%

Table B-4. Selection using video firing model: 75-percent selection ratio

Actual HOPT performance percentile above 25	Predicted performance using video firing model	
	Failure	Success
Success	211	1,148
Failure	234	211

Note: Hit rate = (234 + 1,148)/1,804 = 76.6%
Percentage of correct selections = 1,148/1,359 = 84.5%

Table B-5. Selection using GPA model: 75-percent selection ratio

Actual HOPT performance percentile above 25	Predicted performance using GPA model	
	Failure	Success
Success	205	1,154
Failure	240	205

Note: Hit rate = (240 + 1,154)/1,804 = 77.3%
Percentage of correct selections = 1,154/1,359 = 84.9%

Table B-6. Selection using supervisor rating model: 75-percent selection ratio

Actual HOPT performance percentile above 25	Predicted performance using supervisor model	
	Failure	Success
Success	240	1,119
Failure	205	240

Note: Hit rate = (205 + 1,119)/1,804 = 73.4%
Percentage of correct selections = 1,119/1,359 = 82.3%

Table B-7. Selection using field conduct model: 75-percent selection ratio

Actual HOPT performance percentile above 25	Predicted performance using supervisor model	
	Failure	Success
Success	257	1,102
Failure	188	257

Note: Hit rate = (188 + 1,102)/1,804 = 71.5%
Percentage of correct selections = 1,102/1,359 = 81.1%

APPENDIX C

CROSS-SAMPLE PREDICTION FREQUENCIES FOR THE 25-PERCENT SELECTION RATIO

APPENDIX C

CROSS-SAMPLE PREDICTION FREQUENCIES FOR THE 25-PERCENT SELECTION RATIO

The following tables show the decision outcomes of using regression models based on a HOPT and six surrogates to predict actual hands-on performance of infantry tasks. Each table shows the result of using a 25-percent selection ratio cutoff using a particular model. To build the model, the entire sample of 1,804 was randomly split into two equal groups. Each model was then developed by separately regressing the two groups' criterion (HOPT) or surrogate scores (e.g., job-knowledge test) on the ten enlisted ASVAB subtests (GS, AR, WK, PC, NO, CS, AS, MK, MC, and EI). Regression coefficients from the opposite group were used to cross-validate each model. "Actual success" refers to those who scored in the top 25th percentile in the hands-on criterion, while "actual failure" refers to those who scored in the bottom 75 percent. "Predicted success" means obtaining a score in the top 25th percentile in the predicted criterion (HOPT or surrogate) from the cross-validated regression composite of 10 ASVAB subtests. "Hit rate" is the proportion of all classification decisions made correctly, whereas "percentage of correct selections" is limited to the proportion of successes among those who actually would be selected on the basis of the model.

Table C-1. Selection using HOPT model: 25-percent selection ratio

Actual HOPT performance percentile above 75	Predicted performance using HOPT model	
	Failure	Success
Success	224	259
Failure	1,097	224

Note: Hit rate = (1,097 + 259)/1,804 = 75.2%
Percentage of correct selections = 259/483 = 53.6%

Table C-2. Selection using job-knowledge test model: 25-percent selection ratio

Actual HOPT performance percentile above 75	Predicted performance using core job-knowledge test	
	Failure	Success
Success	237	246
Failure	1,084	237

Note: Hit rate = (1,084 + 246)/1,804 = 73.7%
Percentage of correct selections = 246/483 = 50.9%

Table C-3 Selection using GPA model: 25-percent selection ratio

Actual HOPT performance percentile above 75	Predicted performance using GPA model	
	Failure	Success
Success	237	246
Failure	1,084	237

Note: Hit rate = (1,084 + 246)/1,804 = 73.7%
Percentage of correct selections = 246/483 = 50.9%

Table C-4. Selection using field proficiency model: 25-percent selection ratio

Actual HOPT performance percentile above 75	Predicted performance using field proficiency	
	Failure	Success
Success	252	231
Failure	1,069	252

Note: Hit rate = (1,069 + 231)/1,804 = 72.1%
Percentage of correct selections = 231/483 = 47.8%

Table C-5. Selection using video firing model: 25-percent selection ratio

Actual HOPT performance percentile above 75	Predicted performance using video firing	
	Failure	Success
Success	243	240
Failure	1,078	243

Note: Hit rate = (1,078 + 240)/1,804 = 73.1%
Percentage of correct selections = 240/483 = 49.7%

Table C-6. Selection using supervisor rating model: 25-percent selection ratio

Actual HOPT performance percentile above 75	Predicted performance using supervisory rating	
	Failure	Success
Success	256	227
Failure	1,065	256

Note: Hit rate = (1,065 + 227)/1,804 = 71.6%
Percentage of correct selections = 227/483 = 47.0%

Table C-7. Selection using field conduct model: 25-percent selection ratio

Actual HOPT performance percentile above 75	Predicted performance using field conduct	
	Failure	Success
Success	279	204
Failure	1,042	279

Note: Hit rate = (1,042 + 204)/1,804 = 69.1%
Percentage of correct selections = 204/483 = 42.2%